Proportional-odds models for repeated composite and long ordinal outcome scales

Nick R. Parsons[1]

Warwick Medical School, The University of Warwick,
Coventry, CV4 7AL, U.K.

Abstract

Composite or long ordinal scores, that is scores that have a large number of categories and a natural ordering often resulting from the sum of a number of short ordinal scores, are widely used in many medical studies to assess function or quality of life. Typically these are analysed using unjustified assumptions of normality for the outcome measure, that are unlikely to be even approximately true. Scores of this type are better analysed using methods reserved for more conventional (short) ordinal scores, such as the proportional-odds model. The need for a large number of cut-point parameters, that define the divisions between the score categories, for long ordinal scores in the poroprtional-odds model can be avoided by the inclusion of orthogonal polynomial contrasts. The repeated measures proportional odds logistic regression model is introduced and modifications to the generalized estimating equation methodology used for parameter estimation are described for long ordinal outcomes. Data from a trial assessing two surgical interventions are introduced and briefly described and re-analysed using the new model; inferences from the new analysis are compared to previously published results for the primary outcome measure (hip function at 12 months postoperatively). A simulation study is used to illustrate how this model also has more general application for conventional short ordinal scores, to select amongst competing models of varying complexity for the cut-point parameters.

1 Introduction

Ordinal scores, that have a clear hierarchical ordering, recorded from the same patient, sampling or experimental unit over time are widely used in many medical studies to assess attributes such as pain or to diagnose a condition typically using commonly understood terms such as mild, moderate or severe. It is generally difficult to assess symptoms or conditions of this type in any other way. Typically patients are followed-up or monitored for a fixed period of time and

[1]Email: nick.parsons@warwick.ac.uk

the effect of an intervention assessed based on changes in observed ordinal scores over time. For many more complex characteristics such as function or quality of life, scores from a number of individual ordinal items are often summed to give a composite score. The rationale for measuring a number of items and summing or combining the scores is that it produces a more reliable, sensitive and valid measure than a single item [1]. If the individual items are measuring the same underlying construct (or latent trait) then it seems reasonable to consider the total score to also be a measure of this same construct; although it is difficult to imagine how one might test this hypothesis. Composite scores are sometimes described, to distinguish them from conventional ordinal scores, as long ordinal scores, that is, scores with a large number of categories and a natural ordering.

Clinical scoring systems such as the Oxford Hip Score (OHS) [2] or Multiple Sclerosis Walking Scale (MSWS-12) [3] combine individual items measuring the same construct; hip function for the former and walking ability for the latter score. Although the OHS, for example, is widely used both as a tool to assess treatment effectiveness and also for health service evaluation (e.g. Hospital Episode Statistics for England [4]), there is little guidance for clinicians as to the most appropriate statistical tests to use for comparing treatments or the expected distribution of the scores [2]. There is, however, an acceptance that the score distribution for populations of healthy and sick patients are likely to be skewed [5, 6], due presumably to ceiling and floor effects. Typically scores are considered to be a defendable approximation to an interval scale measuring a latent variable, although in most cases there is little prima facie evidence to indicate that they are anything other than ordinal. Parametric statistical analysis, such as t-tests and linear regression, are usually applied, based on a normal approximation for the distribution of the scores; see Costa et al. [7] for an example of analyses of OHS data. Despite widespread reporting of the results of analyses of long ordinal (composite) scores on an interval scale, it is clearly erroneous to assume that clinically important differences should be equivalent across the whole scale [8]. It may be that parametric analysis of score totals is justified by the Central Limit Theorem for scores within the middle of the score range. However, for many trials a sizeable proportion of participants have recovered or returned to normal health at the trial endpoint, when the definitive analysis is undertaken, and as such are likely to be at the extremes of the score scale. Many composite scores, such as the OHS, are in reality measures of accumulated dysfunctional (or functional) characteristics (e.g. pain, walking ability etc), and it is the accumulation or sum of these attributes that is seen to be clinically important as a kind of overall intensity measure. Good face validity and sensitivity to change (e.g. for OHS [9]) often contributing to their widespread use. A judgement of the relative importance of the attribute items is deliberately not made for the score; recognition that if a weighing did exist, it would almost certainly be different for each person. Scores of this type should not be routinely analysed on the assumption that they are defined on an interval scale, and would be better analysed in a similar manner to conventional ordinal scores. Although we accept in principle that the sum of a number of ordinal variables is itself not necessarily ordinal [10], for many

composite scores, such as the OHS, that measure a single construct it seems a not unreasonable assumption. Particularly given that these composite scores are currently interpreted and information in the data used inefficiently, in much the same manner that ordinal scores are (mis-)treated as if they were continuous variables for the purposes of analysis [11, 12]. Although, whether one could ever truly test whether a composite scale is ordinal is open to question. Choosing to model composite scores, such as the OHS, with many categories (i.e. more than thirty) using ordinal regression models seems to go against convention in this area where normally models are restricted to outcomes with a relatively small number of categories (less than seven). However, there appears to be no strong rationale as to why this should be so, other than the observation that if the distribution is spread over a reasonably large number of categories then it can be assumed that the data were generated from a continuous distribution, on the basis that the scale is likely to be approximately linear [13].

Regarding composite scores observed at a fixed number of occasions (e.g. 6, 12, 24 months) to be ordinal, for the purposes of analysis, suggests that methods developed specifically for analysis of repeated scores of this type may be appropriate. Many approaches to the modelling of repeated ordinal scores have been suggested, and this remains a widely studied statistical problem and an active area of research; see for instance Agresti and Natarajan [14] for a comprehensive review of available models and methods. Without doubt, the most commonly used model for data of this type is the proportional-odds model [15]. The reasons for the widespread use of this model is that the formulation follows naturally from considering an ordinal score to be a continuous (unobserved) variable that is sub-divided into categories to give an evaluation of a quantity that could not be measured directly in any other way. The most widely used method for fitting population-averaged proportional-odds models to repeated ordinal scores uses the generalized estimating equation (GEE) method originally proposed by Liang and Zeger [16]. There are various options for setting-up the model and estimating parameters [17, 18, 19]; the method discussed here is the so-called repeated measures proportional odds logistic regression model [20] available in the R [21] function `repolr`. For the proportional-odds model, the cut-point parameters that define divisions between the ordinal score categories are estimated together with regression parameters that characterise the treatment effects. The cut-point parameters are often regarded as nuisance parameters that cannot easily be interpreted. For long ordinal scores, with a large number of categories, this presents a problem, as many cut-point parameters would need to be estimated with presumably poor precision and likely convergence problems that are often particularly associated with models for repeated ordinals scores [22, 23]. However, it seems that almost by definition as the number of categories increases there will generally be less interest in individual cut-point parameter estimates and more interest in the shape of the response, particularly at the extremes of the score scale where deviations from linearity are likely to be important. This suggests that modelling the cut-point parameters may prove to be practical in this setting. Orthogonal polynomials, that partition the variation due to the cut-points into single degree of freedom orthogonal polyno-

mial contrasts, provide a convenient model for the cut-point parameters. These polynomial models can be incorporated into existing GEE models for repeated ordinals score to allow modelling of composite scores and thereby address what appears to be the only obstacle to the widespread use of these models for scores of this type. Others have suggested modelling cut-point parameters using generalized logistic and non-parametric functions in proportional-odds models for continuous ordinal scores derived from visual analogue scales in a Bayesian setting [24], but this is the first paper to develop models for long truly ordinal scores.

Section 2 develops a cut-point parameter model and describes estimation and interpretation. Example data are introduced and analyses described in Section 3, a simulation study is undertaken in Section 4 to explore the behaviour of the cut-point model and the wider application of the model is discussed in Section 5.

2 Model

2.1 Proportional-odds

Let N experimental units be scored at each of T time points using an ordinal scale with K categories, where the scores are made on an integer-valued scale from 1 to K , where K represents the optimum score. The score on the i^{th} experimental unit at the t^{th} time is represented by Y_{it} and the vector of scores for the i^{th} experimental unit over the set of T time points by $Y_i' = (Y_{i1}, Y_{i2}, \ldots, Y_{iT})$. A multivariate vector of measured variables, x_{it}, is also observed on each experimental unit at each time point t. The probability $\mu_{itk} = P(Y_{it} \leq k)$ for ordinal score Y_{it} can be related to the measured variables x_{it} by a proportional-odds model based on cumulative logits [25]

$$\log\left(\frac{\mu_{itk}}{1-\mu_{itk}}\right) = \gamma_K + x_{it}'\beta. \tag{1}$$

The cut-points $\gamma_K = (\gamma_1, \gamma_2, \ldots, \gamma_{K-1})$, where $-\infty < \gamma_1 < \gamma_2 < \cdots < \gamma_{K-1} < \infty$, define the divisions between the ordinal score categories on the cumulative logit scale and effectively transform the ordinal scale to a continuous scale based on the linear predictor $x_{it}'\beta$. Model parameters γ_K and β can be estimated using modifications to the well known method of generalized estimating equations (GEE), originally proposed by Liang and Zeger [16]; see for instance repeated measures proportional odds logistic regression described in detail by Parsons et al. [20] and available as function library `repolr` in R [21].

Before developing a new approach to modelling the cut-point parameters γ_K it is informative to review how model (1) has previously been modified to accommodate 'continuous' ordinal outcomes, albeit using a very different methodology for parameter estimation. Manuguerra and Heller [24] discuss a cumulative logistic ordinal model for continuous response variables $\log(\nu/(1-\nu)) = g(\nu) + x'\beta$ where the function $g(\nu)$ is a continuous analogue of the (discrete)

cut-point parameters $-\infty < \gamma_1 < \gamma_2 < \cdots < \gamma_{K-1} < \infty$ in the conventional proportional-odds model The differentiable and increasing function $g(\nu)$ maps the continuous ordinal score ν, on the scale $(0,1)$, to a notionally latent variable on the scale $(-\infty,\infty)$; for instance a generalized logistic function $g(\nu) = M + B^{-1}\log\left(T\nu^T/\left(1-\nu^T\right)\right)$, with parameters M (intercept), B (slope) and T (symmetry). This specification of $g(\nu)$ is problematic if the continuous ordinal score ν is defined on a closed (rather than open) interval as, if the extremes of ν are observed, the model is not identifiable. This problem is overcome by converting observations made at the extremes to arbitrarily small (0.001) and large (0.999) values of ν within the open interval. This approach is feasible for the applications Manuguerra and Heller describe, where essentially continuous observations made on a visual acuity scale are converted to ordinal scores. However, it would be nonsensical and simply unacceptable to re-code data in an analogous manner for the long ordinal outcome scales discussed here, particularly as behaviour at the extremes of the scale is of prime interest. A much more natural modification of model (1), in the setting of linear models, is to use an orthogonal polynomial decomposition of the ordered but unstructured cut-point parameters γ_K. This retains the interpretation of γ_K, as divisions between ordinal score categories on the cumulative logit (latent) scale, whilst at the same time imposing constraints on the relative sizes of the score categories; from equal spacing for the most simple model to totally unstructured for the most complex possible model. Splines could also potentially be used as basis functions within the conventional inferential regression framework, and may in principle offer more parsimonious fits for some types of data. However, they lack the simple interpretation provided by orthogonal polynomials and implementation would be complicated by knot selection [26]; unless more computationally involved methods such as penalized regression splines [27, 24] were used, an unnecessary level of complexity for inferences on what are effectively ancillary parameters.

2.2 Cut-points γ_K

As an alternative to estimating $K-1$ cut-point parameters in the conventional manner, the cut-points can be modelled by orthogonal polynomials

$$\gamma_K = \beta_0 F(o_K) \tag{2}$$

for ordinal variables o_K , where $o_K \in (1,2,3,\ldots,K)$, and regression parameters $\beta_0 = \left(\beta_{00},\beta_{01},\beta_{02},\ldots,\beta_{0(K-1)}\right)$. The complete $(K-1)\times(K-1)$ matrix of polynomials $F'(o_K) = [f_0(o_K), f_1(o_K), f_2(o_K), \ldots, f_{K-2}(o_K)]$, where $f_j(o_K)$ is a polynomial of degree j in o_k for normal orthogonal vectors $f_0(o_K), f_1(o_K), f_2(o_K), \ldots, f_{K-2}(o_K)$, is such that

$$\sum_{j=1}^{K-2} f_j(o_K) f'_{j'}(o_K) = \left\{ \begin{array}{l} 1 \text{ if } j = j' \\ 0 \text{ if } j \neq j' \end{array} \right\}. \tag{3}$$

For example, for the five category ordinal scale $o_5 \in (1,2,3,4,5)$, the mean,

linear, quadratic and cubic components defining the divisions between the ordinal score categories are as follows;
$f_0(o_5) = (1,1,1,1)$, $f_1(o_5) = (-3,-1,1,3)/\sqrt{20}$, $f_2(o_5) = (1,-1,-1,1)/2$ and $f_3(o_5) = (-1,3,-3,1)/\sqrt{20}$, where $F(o_5)F'(o_5) = \mathrm{I}_4$ as required; I_4 is the 4×4 identity matrix. Clearly the complete matrix of orthogonal polynomials simply provides an alternative parameterization for the cut-point parameters to the conventional proportional odds model where, for the chosen example, rather than including polynomial contrasts, setting $F(o_5) = \mathrm{I}_4$ gives $\gamma_1 = \beta_{00}$, $\gamma_2 = \beta_{01}$, $\gamma_3 = \beta_{02}$ and $\gamma_4 = \beta_{03}$. However, by using a selection of lower order polynomial terms model (2) allows a simpler model for γ_K to be fitted than would be possible with the conventional parameterization. This is a particularly useful property for ordinal scales where K is large (>10) or for instance where a specific form for the cut-points parameters is preferred, e.g. where it is desirable that divisions between score categories increase uniformly towards the extremes of the scale. Also, as model (2) does not require a full specification of $K-1$ cut-point parameters, it naturally handles data where one or more ordinal score categories are not recorded, by appropriate modification of the orthogonal polynomials for unequal spacing [28].

2.3 Estimation

Proportional-odds models incorporating polynomial terms for γ_K can be fitted, after modification of the complete data matrix X_i , using the iterative scheme suggested by Parsons et al. [23] and implemented in the R [21] function `repolr` [20]. Using the notation of Parsons et al. [23] for S explanatory variables x_{it} observed at each of the T time points and $T \times S$ matrix of explanatory variates for each sampling unit $i = 1, \ldots, N$ given by $X_{0i} = [x_{i1}, x_{i2}, \ldots, x_{iT}]'$, the complete data matrix including the cut-points and the explanatory variables for the i^{th} sampling unit is $X_i = [1_T \otimes \mathrm{I}_{K-1}, X_{0i} \otimes 1_{K-1}]$. Here, 1_T and 1_{K-1} are T dimensional and $K-1$ dimensional vectors of unit elements, respectively, and I_{K-1} is the $K-1$ dimensional identity matrix. For the new polynomial cut-point model, the matrix I_{K-1} that identifies the individual cut-point parameters, can be replaced by the $(K-1)\times(J+1)$ matrix $F_J(o_K)$ of orthogonal polynomials of degree $J \leq K-2$. For the example five category ordinal scale $o_5 \in (1,2,3,4,5)$, a model with mean, linear and quadratic terms can be implemented with

$$F_2(o_5) = \tfrac{1}{2}\begin{bmatrix} 1 & -3/\sqrt{5} & 1 \\ 1 & -1/\sqrt{5} & -1 \\ 1 & 1/\sqrt{5} & -1 \\ 1 & 3/\sqrt{5} & 1 \end{bmatrix}$$

and the $T(K-1) \times (2+1+S)$ dimensional complete data matrix for the i^{th} sampling unit $X_i = [1_T \otimes F_2(o_5), X_{0i} \otimes 1_{K-1}]$. Model parameters are estimated by solving an estimating equation $Q(\beta, \beta_0; \alpha) = 0$ for β and β_0, where the association parameter α is estimated by minimising the logarithm of the determinant of the covariance matrix of the regression parameters at each step of

the fitting algorithm, for a range of correlation models [20]. Code to implement the polynomial cut-point models discussed here is available on request from the author, and will also be incorporated in future releases of `repolr`.

2.4 Parameter interpretation

For the conventional model, where we set $F(o_K) = \mathrm{I}_{K-1}$ in model (2), the parameters β_0 (where $\gamma_K = \beta_0$) are identified as the divisions between the ordinal score categories on the cumulative logit scale. More generally parameter estimates $\tilde{\beta}_0$ represent polynomial effects that are not constrained in the manner of γ_K, but can simply be transformed to give conventional cut-point estimates $\tilde{\gamma_K} = \tilde{\beta}_0 F(o_K)$. Model selection proceeds by calculating the quasilikelihood under the independence model information criterion (QIC), which is given by $\mathrm{QIC} = -2Q_{\mathrm{I}} + 2p$, where Q_{I} is the quasilikelihood for the independence model and p (penalty term) is the number of model parameters [29]. Models of increasing polynomial complexity for γ_K can be fitted and the model with the smallest QIC criterion measure selected.

3 Example

3.1 Data

The Warwick Arthroplasty Trial [7] is a randomized controlled trial that compared hip function in patients after surgery using one of two procedures, resurfacing arthroplasty (RSA) and total hip arthroplasty (THA) [30]. A range of outcome measures were reported at 6 weeks (t_0), 3 months (t_1), 6 months (t_2) and 12 months (t_3) after surgery for 126 patients (60 in RSA and 66 in THA). Foremost amongst these measures was the Oxford hip score (OHS). The OHS is a patient-reported measure comprising of 12 ordinal score items, on a scale from 0 to 4, that are summed for a patient to give an overall score that quantifies hip function on a scale from 0 to 48, where 0 indicates very poor hip function and 48 excellent hip function. A difference of 5 points in the OHS at 12 months between treatment groups was considered to be clinically important and this was used to formally power the study. The primary analysis of the OHS, reported previously, compared mean scores between groups using a t-test based on a normal approximation for the distribution of the score at 12 months, using an intention-to-treat analysis of complete data [7].

3.2 Analysis

Figure 1 shows patient counts of OHS at each assessment occasion. There is a clear trend for increasing OHS with successive assessment occasions post surgery, with many trial participants returning to excellent function (>42) [31] by 12 months. Costa et al. [7] report mean scores and 95% confidence intervals at the trial endpoint of 12 months for the RSA and THA groups of 40.4 (37.9, 42.9) and 38.2 (35.3, 41.0), and a p-value of 0.242 from a t-test ($\mathrm{t} = 1.176$, d.f. =

118). Adjusting for age and gender in a regression analysis made no qualitative difference to the estimated treatment effect and the authors concluded that although there was no evidence for a significant treatment effect in these data, the estimated confidence interval (-1.52, 5.98) for the treatment effect (RSA versus THA; 2.23) suggested that a clinically important effect could not be ruled out; i.e. the minimum clinically important difference of 5 points was contained within the 95% CI. Figure 1 shows that the OHS distribution was far from normal by the trial endpoint. Consequently, given the poor adherence to the distributional assumptions required for the reported analyses, is it sure that there really was no significant treatment effect? And if this is indeed the case, can stronger evidence be provided to support the analysis undertaken by Costa et al. [7] and rule out a difference in OHS larger than 5 points.

Estimated polynomial cut-point parameters β_0 and QIC values are shown in Table 1 for proportional-odds models for the OHS, including treatment by time interaction terms, a first order autoregressive correlation structure for the association parameter [20] and using linear, quadratic, cubic and quartic orthogonal polynomial models for the cut-points. QIC values for higher order polynomials suggested that there was no improvement for more complex models ($\mathrm{Q}_{quintic} = 13021.94$, $\mathrm{Q}_{sextic} = 13024.25$ and $\mathrm{Q}_{septic} = 13025.58$). Forty four of the available 48 score categories were observed for this population; therefore it was necessary to make appropriate modification of the orthogonal polynomials for the unequal spacing of categories. Transforming the polynomial cut-point parameters (Table 1) to the more conventional cut-points parameters γ_K using equation 2 gives the curves shown in Figure 2. The QIC suggests that a quartic model for the cut-point parameters provides the most parsimonious model, with important departures from linearity towards the lower end of the score scale (Figure 2). The first order autoregressive correlation coefficient was small with a point estimate of 0.14 for the quartic model. Regression parameter estimates for the quartic model, shown in Table 2, indicate significant temporal effects, evidenced by the overall improvement in scores as the trial participants recovered function after surgery, but a lack of significant treatment effect. Interpretation of results from ordinal regression analysis can be aided by presentation on the original score scale [32], using estimated score category probabilities from equation 1. This can be combined with nonparametric cluster bootstrap resampling [33], using 1000 samples with replacement from trial participants rather than from individual data points, to give an estimate for the treatment group difference and 95% confidence interval of 0.66 (-1.82, 3.15). With the conclusion that a treatment group difference of 5 points seems to be much less likely than was suggested by the analysis of Costa et al. [7].

4 Simulation study

The methodology proposed here, using polynomial terms to model the cut-point parameters, also has application for analysis of short ordinal scores. Therefore a simulation study was undertaken to explore the behaviour and properties of

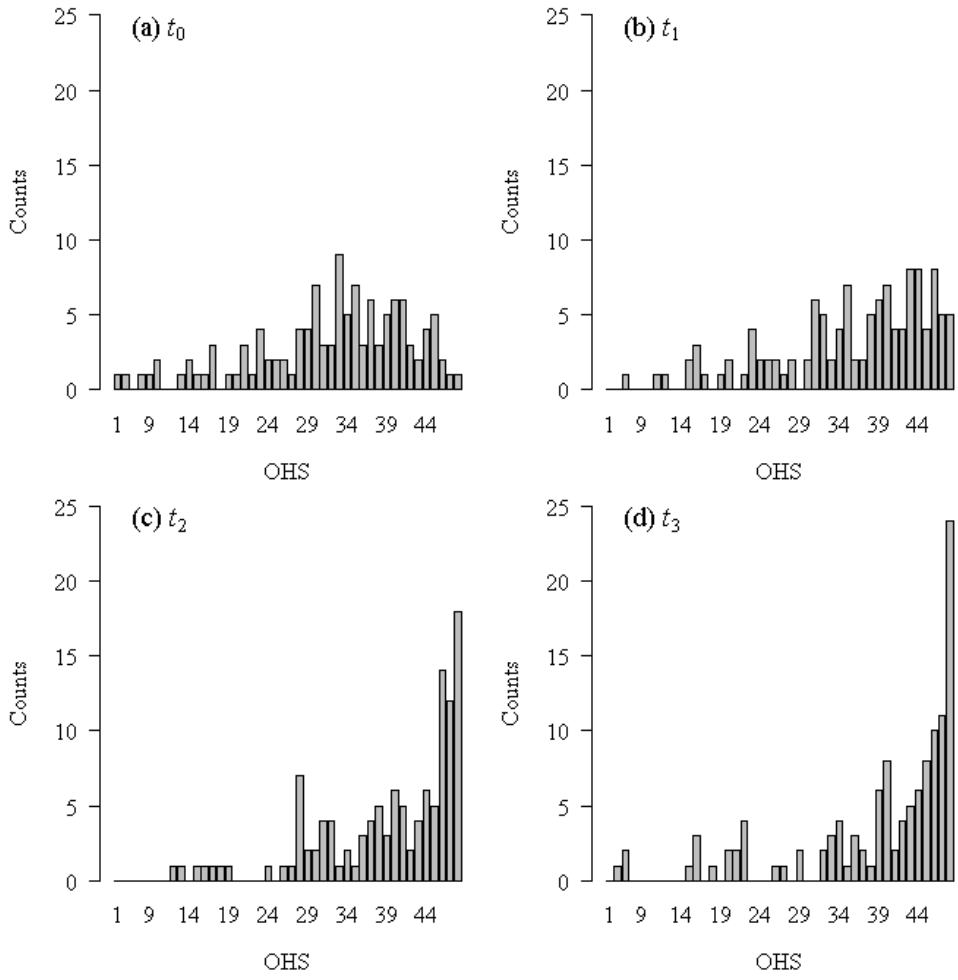

Figure 1: Counts of OHS scores by category and assessment occasion; (a) 6 weeks (t_0), (b) 3 months (t_1), (c) 6 months (t_2) and (d) 12 months (t_3).

Table 1: Estimated polynomial cut-point parameters β_0, with standard errors, and QIC values for 1^{st}, 2^{nd}, 3^{rd} and 4^{th} order polynomial models.

	Order							
	1^{st}		2^{nd}		3^{rd}		4^{th}	
	Est.	s.e.	Est.	s.e.	Est.	s.e.	Est.	s.e.
β_{00}	-1.41	0.26	-1.33	0.25	-1.39	0.26	-1.42	0.28
β_{01}	258.24	20.43	245.76	20.69	264.00	25.52	271.38	30.75
β_{02}			7.32	8.42	-17.79	16.42	-26.65	23.93
β_{03}					16.20	8.44	25.41	17.31
β_{04}							-5.19	7.39
QIC	13028.77		13035.12		13020.01		13019.85	

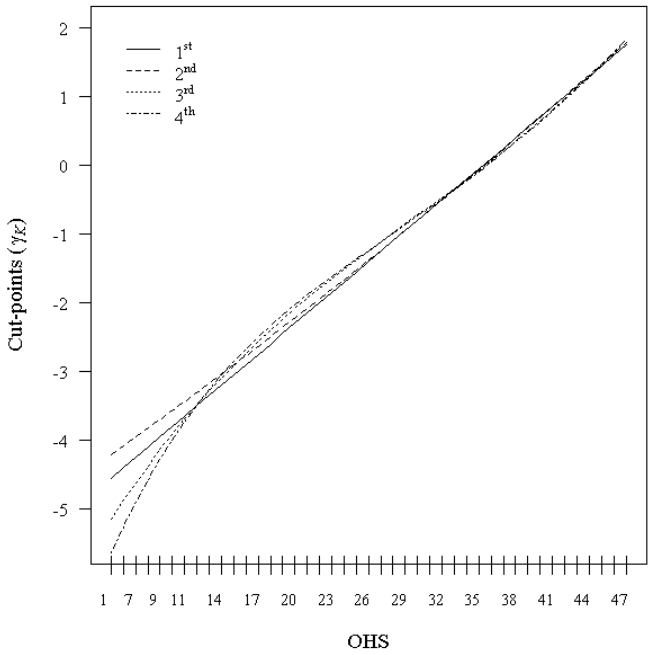

Figure 2: Cut-point estimates from OHS data for polynomial models of 1^{st}, 2^{nd}, 3^{rd} and 4^{th} order.

Table 2: Estimated regression parameters β, with standard errors, for the 4^{th} order polynomial model; where $\beta_{t_0} = 0$ and $\beta_{RSA} = 0$.

	Est.	s.e.	Z
β_{t_1}	-0.79	0.17	-4.60
β_{t_2}	-1.55	0.22	-6.97
β_{t_3}	-1.88	0.26	-7.33
β_{THA}	-0.29	0.30	-0.98
$\beta_{t_1} \times \beta_{THA}$	0.21	0.23	0.88
$\beta_{t_2} \times \beta_{THA}$	0.31	0.31	1.02
$\beta_{t_3} \times \beta_{THA}$	0.64	0.34	1.88

the proposed polynomial cut-point model compared to the usual fully specified proportional-odds cut-point model, in a less complex setting than the example of Section 3 using ordinal outcomes with 10 score categories. This simpler setting also allows the process of model selection described in the data example to be understood more fully, particularly potential biases that may arise when estimating model regression parameters. Ordinal data with 10 score categories were simulated using the `ordsample` function available in the `GenOrd` library [34] in `R` [21] for 250 subjects, divided into two equally sized groups, at two correlated timepoints (correlation parameter 0.6) and with randomly chosen cut-points in two contrasting settings; (i) there was no difference in scores between groups ($\beta = 0$) and (ii) scores were higher in one group than the other ($\beta = 0.8$). The conventional proportional-odds model, together with models that used linear, cubic, quintic and septic orthogonal polynomial models for the cut-points, were fitted to each of 1000 data simulations, and model parameters estimated. Mean cut-point parameter estimates are shown in Figure 3. Mean values of QIC suggested an improvement in model fit for setting (i) with increasing complexity of polynomial models ($\text{Q}_{linear} = 5089.02$, $\text{Q}_{cubic} = 5081.17$, $\text{Q}_{quintic} = 5066.54$ and $\text{Q}_{septic} = 5062.32$), whereas for setting (ii) there was scant evidence for an improvement in fit for increasing complexity ($\text{Q}_{linear} = 4704.21$, $\text{Q}_{cubic} = 4698.95$, $\text{Q}_{quintic} = 4699.90$ and $\text{Q}_{septic} = 4702.67$). This can be seen in Figure 3, where for setting (i) the higher order polynomial curves provide closer fits to the conventional proportional-odds model estimates of the cut-points. In setting (ii), the cut-point spacing model is by chance (due to the random selection of cut-points) much more linear, and as such higher order polynomial models provide little or no improvement in fit. Estimates of the regression parameter for each polynomial model ($\beta_{1^{st}}$, $\beta_{3^{rd}}$, $\beta_{5^{th}}$ and $\beta_{7^{th}}$) were compared to the those from the conventional proportional-odds model (β_{po}). Figure 4 shows that, as expected, for both settings, there was less scatter with increasing cut-point model complexity; regression parameter estimates from the polynomial models tended towards estimates from the conventional proportional-odds model, with increasing cut-point model complexity. In setting (i) there was no evidence of bias in the parameter estimates with increasing complexity; that is, differences between regression parameter estimates $\beta_{po} - \beta_{j^{th}}$ were distributed approximately sym-

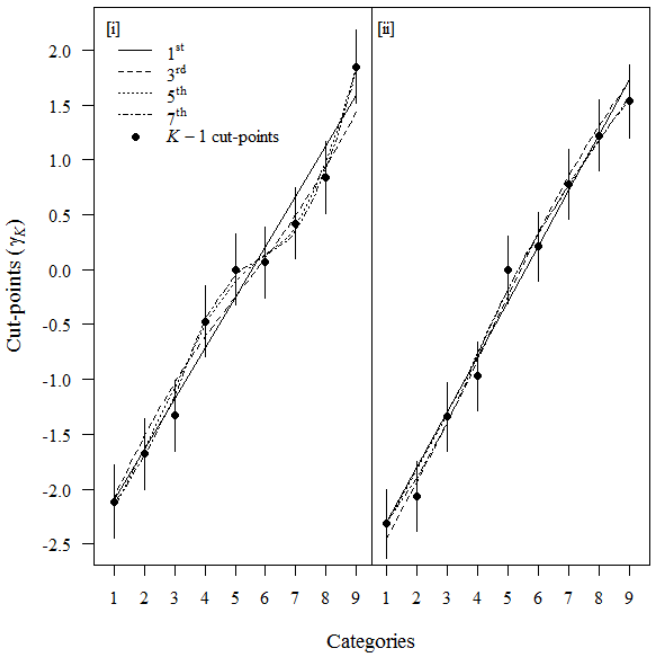

Figure 3: Cut-point estimates for polynomial models of 1^{st}, 3^{rd}, 5^{th} and 7^{th} order and conventional $K-1$ cut-point model (with 95% confidence intervals) for simulation settings [i] and [ii].

metrically about zero, with interquartile ranges of (-0.0331, 0.0367), (-0.0207, 0.0218), (-0.0103, 0.0104) and (-0.0038, 0.0035) for linear, cubic, quintic and septic cut-point models. Whereas for setting (ii), where there was little evidence from QIC for models more complex than linear, regression parameter estimates tended to be larger for higher order polynomial models, albeit by a relatively small amount, than those from the fully specified proportional-odds cut-point model; interquartile ranges were (-0.0148, 0.0328), (-0.0085, 0.0223), (-0.0006, 0.0177) and (-0.0006, 0.0162) for linear, cubic, quintic and septic cut-point models. In summary, there is inevitably some loss of precision in regression parameter estimates when using the simpler polynomial cut-point model (with fewer parameters), relative to the usual fully specified proportional-odds cut-point model. When appropriate modelling strategies are used, that is cut-point model complexity is informed by QIC, then parameter estimates were unbiased. However, care must be taken when attempting to fit unnecessarily complex polynomial models where there is no supporting evidence, as this may lead to some (albeit small) bias in regression parameter estimates.

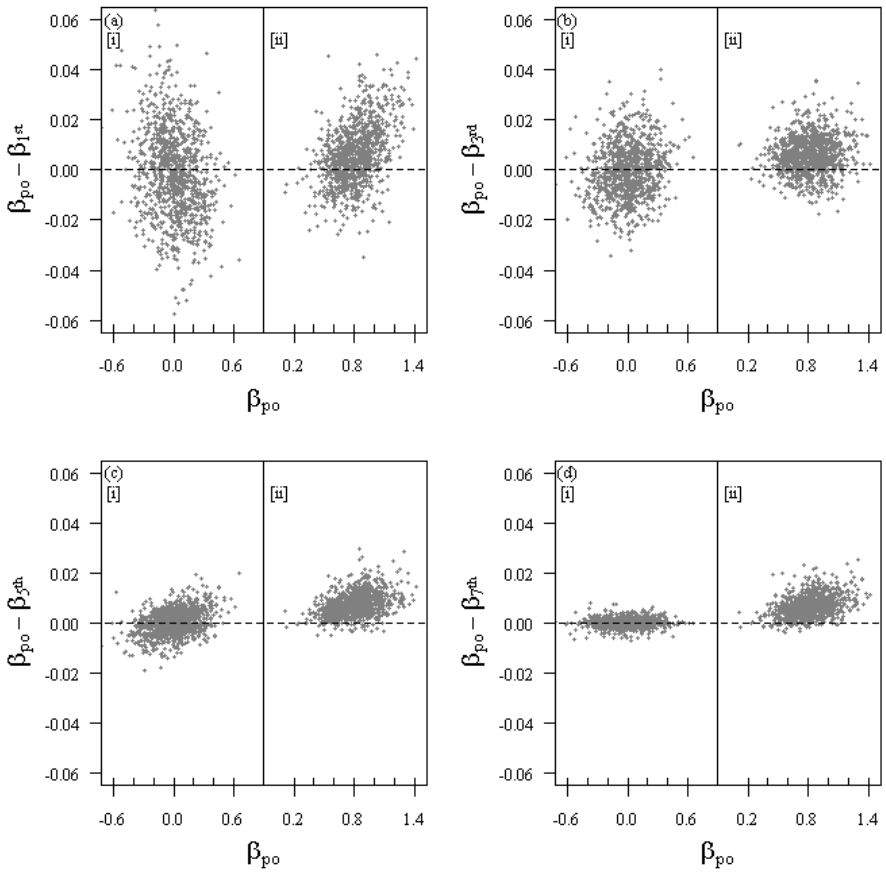

Figure 4: Scatter plots of $\beta_{po} - \beta_{j^{th}}$ versus β_{po} for linear (a), cubic (b), quintic (c) and septic (d) polynomials for simulation settings [i] and [ii].

5 Discussion

Scoring systems for complex characteristics such as function or quality of life are widely reported in medicine and social science. Typically if there are a small number of categories the scores are assumed to be ordinal and appropriate methods for analysing data of this type are employed. However, for long ordinal scores (with many categories) or composite scores, consisting of the sums of many shorter scores, analysis is usually based on assumptions of normality for the score distributions (e.g. t-tests, linear regression). This assumption is unlikely to be even approximately true in for instance many clinical scenarios where the outcome for most patients will be a return to normal function at the study endpoint (the extremum of the scale). The selection of the method of analysis based on the number of categories in the score scale seems arbitrary and more to do with the perceived shortcomings of current ordinal methods for anything more than a small number of categories (<10) than a belief that scores with many categories can really be assumed to come from a continuous distribution.

For the widely used proportional-odds model, a simple modification to the cut-point model allows the number of estimated parameters to be reduced drastically for ordinal scores with many categories, by constraining the form of the spacing model that defines the divisions between the score categories to be a polynomial of degree $J \leq K-1$. Experience suggests that cubic or quintic polynomials are usually of sufficient complexity to characterise the spacing model between categories. For long ordinal scores, which for instance may have 50 categories, this reduces the number of cut-point parameters that need to be estimated from 49 to no more than 4 or 6. By reducing the number of model parameters, the methodology proposed here overcomes potential problems associated with attempting to estimate cut-point parameters for saturated cut-point models for repeated ordinals scores [23]. Also, as fewer cut-point parameters are estimated they are less likely to simply be dismissed as nuisance parameters, and as their link to the category spacing model is made more explicit (e.g. linear, cubic term), their interpretation is much easier. More generally, constraining the form of the spacing model that defines the divisions between the core categories is probably a desirable property that should also be considered for shorter ordinal scores (Section 4). The simulation study showed that polynomial models are good alternatives to conventional fully specified (saturated) cut-point models for short ordinal outcomes, and thus by extension it is expected that the desirable properties reported in Section 4 are also true for long ordinal scores, where fitting the fully specified cut-point model would be at best problematic and generally not possible.

Inferences from analyses are always strongly dependent on modelling assumptions, and this was the case for the example data in Section 3.1. Fitting the proportional-odds model to the OHS data from the WAT suggested that the interpretation of Costa et al. [7], based on assumed normality for the score distribution, were unduly cautious. The analysis presented here suggests that the true treatment difference is considerably smaller than that reported previously

(0.66 rather than 2.23) and the (95%) confidence interval for the treatment effect is such that a difference of 5 points can be ruled out as very unlikely. It may be argued that composite scores such as the OHS should be differentiated from other ordinal scores, as there is clearly no guarantee that the sum of a number of ordinal variables is itself ordinal. However, for many composite scores the individual components are rarely if ever reported or assessed separately and the overall scores are treated and interpreted exactly as if they were ordinal. Given that this is the case the approach suggested here for analysis seems appropriate and therefore more efficient and sensitive than that offered by models based on normal approximations.

References

[1] Cox D, Fitzpatrick R, Fletcher A, Gore S, Spiegelhalter D, Jones D. Quality-of-life assessment: can we keep it simple? *Journal of the Royal Statistical Society Series A* 1992; **155**:353–393.

[2] Dawson J, Fitzpatrick R, Carr A, Murray D. Development and validation of a questionnaire to assess patients' perceptions in relation to total hip replacement surgery. *Journal of Bone and Joint Surgery (British)* 1996; **78**:185–190.

[3] Hobart J. Rating scales for neurologists. *Journal of Neurology Neurosurgery and Psychiatry* 2003; **74**:iv22–iv26.

[4] Hospital episode statistics for england 2012. URL http://www.hesonline.nhs.uk.

[5] Hunsacker F, Cioffi D, Amadio P, Wright J, Caughlin B. The american academy of orthopaedic surgeons outcomes instruments: normative values from the general population. *Journal of Bone and Joint Surgery (American)* 2002; **84**:208–215.

[6] Murray D, Fitzpatrick R, Rogers R, Pandit H, Beard D, Carr A, Dawson J. The use of the oxford hip and knee scores. *Journal of Bone and Joint Surgery (British)* 2007; **89**:1010–1014.

[7] Costa M, Achten J, Parsons N, Edlin R, Foguet P, Prakash U, Griffin D. A randomised controlled trial of total hip arthroplasty versus resurfacing arthroplasty in the treatment of young patients with arthritis of the hip joint. *BMJ* 2012; **344**:e2147.

[8] Horton M, Tennant A. Patient reported outcomes: misinference from ordinal scales? *Trials* 2011; **12**:A65.

[9] Dawson J, Fitzpatrick R, Frost S, Gundle R, McLardy-Smith P, Murray D. Evidence for the validity of a patient-based instrument for assessment of outcome after revision hip replacement. *Journal of Bone and Joint Surgery (British)* 2001; **83**:1125–1129.

[10] Forrest M, Andersen B. Ordinal scale and statistics in medical research. *BMJ* 1986; **292**:537–538.

[11] Lavalley M, Felson D. Statistical presentation and analysis of ordered categorical outcome data in rheumatology journals. *Arthritis and Rheumatism* 2002; **47**:255–259.

[12] Jakobsson U. Statistical presentation and analysis of ordinal data in nursing research. *Scandinavian Journal of Caring Sciences* 2004; **18**:437–440.

[13] Walters S, Campbell M, Lall R. Design and analysis of trials with quality of life as an outcome: a practical guide. *Journal of Biopharmaceutical Statistics* 2001; **11**:155–176.

[14] Agresti A, Natarajan R. Modeling clustered ordered categorical data: a survey. *International Statistical Review* 2001; **69**:345–371.

[15] McCullagh P. Regression models for ordinal data. *Journal of the Royal Statistical Society Series B* 1980; **42**:109–142.

[16] Liang K, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.

[17] Clayton D. Repeated ordinal measurements: a generalized estimating equations approach. Technical Report, MRC Biostatistics Unit, Cambridge, UK 1992.

[18] Lipsitz S, Kim K, Zhao L. Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine* 1994; **13**:1149–1163.

[19] Kenward M, Lesaffre E, Molenberghs G. An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* 1994; **50**:945–954.

[20] Parsons N, Costa M, Achten J, Stallard N. Repeated measures proportional odds logistic regression analysis of ordinal score data in the statistical software package r. *Computational Statistics & Data Analysis* 2009; **53**:632–641.

[21] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2011. URL http://www.R-project.org, ISBN 3-900051-07-0.

[22] Lipsitz S, Fitzmaurice G, Endel J, Laird N. Performance of generalized estimating equations in practical situations. *Biometrics* 1994; **50**:270–278.

[23] Parsons N, Edmondson R, Gilmour S. A generalized estimating equation method for fitting autocorrelated ordinal score data with an application in horticultural research. *Journal of the Royal Statistical Society Series C* 2006; **55**:507–524.

[24] Manuguerra M, Heller G. Ordinal regression models for continuous scales. *International Journal of Biostatistics* 2010; **6**:14.

[25] McCullagh P, Nelder J. *Generalized Linear Models*. Chapman & Hall, 1989.

[26] Hastie T, Tibshirani R. *Generalized Additive Models*. Chapman & Hall, 1990.

[27] Green P, Silverman B. *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, 1994.

[28] Robson D. A simple method for constructing orthogonal polynomials when the independent variable is unequally spaced. *Biometrics* 1959; **15**:187–191.

[29] Hardin J, Hilbe J. *Generalized Estimating Equations*. Chapman & Hall, 2002.

[30] Achten J, Parsons N, Edlin R, Griffin D, Costa M. A randomised controlled trial of total hip arthroplasty versus resurfacing arthroplasty in the treatment of young patients with arthritis of the hip joint. *BMC Musculoskeletal Disorders* 2010; **11**:8.

[31] Kalairajah Y, K KA, Hulme C, Molloy S, Drabu K. Health outcome measures in the evaluation of total hip arthroplastiesa comparison between the harris hip score and the oxford hip score. *Journal of Arthroplasty* 2005; **20**:1037–1041.

[32] Hannah M, Quigley P. Presentation of ordinal regression analysis on the original scale. *Biometrics* 1996; **52**:771–775.

[33] Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[34] Ferrari P, Barbiero A. Simulating ordinal data. *Multivariate Behavioral Research* 2012; **47**:566–589.